Alignement d'unités textuelles de taille variable

Emmanuel Giguet, Marianna Apidianaki

ALIGNEMENT D'UNITÉS TEXTUELLES DE TAILLE VARIABLE

Emmanuel Giguet (1), Marianna Apidianaki (2)
GREYC – CNRS UMR 6072– Université de Caen – Bd du Maréchal Juin – F14032 Caen Cedex Emmanuel.Giguet@info.unicaen.fr (2) Lattice – CNRS / ENS / Université Paris 7 – Denis Diderot Ecole Normale Supérieure, 1 rue Maurice Arnoux, F-92120 Montrouge Marianna.Apidianaki@linguist.jussieu.fr

1. INTRODUCTION

1.1. Traitement automatique de corpus bilingues et multilingues

Le traitement de corpus bilingues et multilingues parallèles, constitue un champ d'investigation très animé dans le domaine du traitement automatique des langues. Ces corpus sont constitués d'un ensemble de textes et de leurs traductions dans une ou plusieurs langues. Les informations linguistiques et traductionnelles qui peuvent être mises à jour par l'investigation et l'analyse de ces corpus les rendent une ressource importante tant pour les traducteurs, les lexicographes et les terminologues que pour les linguistes qui utilisent des méthodes empiriques pour l'étude contrastive des langues. Une autre raison qui a nourri l'intérêt envers ce type de ressources est qu'elles constituent le noyau des Systèmes de Traduction Basés sur l'Exemple et des Systèmes de Mémoires de Traduction.

Aujourd'hui, ces corpus sont plus aisément accessibles. De nouveaux sont proposés régulièrement par la communauté, que ce soit pour de nouvelles paires de langues que pour des thématiques spécifiques. Ils mettent à l'épreuve les modèles tant dans leur capacité à appréhender des langues variées qu'à s'adapter à des lexiques variés. L'augmentation de la taille de ces corpus pousse quant à elle vers la recherche de modèles opérationnels efficaces ne demandant que peu de supervision, tout en produisant de bons résultats, que ce soit du point de vue quantitatif ou bien qualitatif, et cela, dans le but de faciliter la description, l'investigation et l'analyse des informations contenues dans les textes.

1.2. Méthodes d'alignement de corpus parallèles

Un sujet « épineux » mais crucial pour toute la recherche tournant autour des corpus bilingues et multilingues est l'alignement, c'est-à-dire la localisation et l'extraction des relations de traduction existant entre les diverses unités constitutives des textes parallèles. Les méthodes d'alignement actuelles se situent sur un continuum allant des méthodes purement statistiques à celles intégrant des connaissances linguistiques détaillées. La nature des connaissances n'est en fait pas la seule manière de classer ces méthodes. Un angle tout aussi pertinent pour les étudier concerne la granularité des alignements proposés : la méthode propose-t-elle des équivalences de paragraphes, de phrases, de propositions, de termes, ou bien de mots, un choix qui dépend en fait des applications envisagées.

Pour l'alignement à gros grain (paragraphe, phrase), les principaux indices utilisés sont la longueur relative des segments, la présence de formes invariantes (certaines

ponctuations, certains symboles, nombres, sigles, etc.) et la mise en forme. Certaines méthodes vont dès ce stade introduire des dictionnaires bilingues. Ainsi, au « grain » phrase, Gale & Church (1991b), Brown *et al.* (1991) et Kay & Röscheisen (1993) proposent des méthodes d'alignement statistiques ; Simard *et al.* (1992) et Papageorgiou *et al.* (1994) ont, eux, recours aussi à des informations linguistiques.

À granularité fine, à savoir sous-phrastique, les méthodes statistiques se sont rapidement vues combiner à des descriptions linguistiques enrichies *via*, par exemple, un étiquetage morpho-syntaxique (tagging) ou une segmentation en groupes minimaux (chunking). Ainsi, on trouve des méthodes statistiques chez Gale & Church (1991a), Kitamura & Matsumoto (1995) et d'autres combinant des informations linguistiques (Tiedemann, 1993 ; Boutsis & Piperidis, 1996 ; Piperidis *et al.*, 1997). On notera également l'existence de méthodes proposant des alignements à un grain intermédiaire entre la phrase et le mot isolé (Smadja, 1992 ; Smadja *et al.*, 1996 ; Kupiec, 1993 ; Kumano & Hirakawa, 1994 ; Boutsis & Piperidis, 1998).

1.3. Limite des méthodes d'alignement

Le problème de l'alignement est par définition celui de la localisation et de la délimitation précise des segments à mettre en correspondance entre les langues. Même si l'alignement au grain paragraphe ou phrase semble plus accessible qu'un alignement sous-phrastique, on y rencontre aussi des difficultés. Une des faiblesses tient entre autre au fait que les méthodes d'alignement misent sur le parallélisme de la structure des documents et ne sont par conséquent que très peu tolérantes aux variations dispositionnelles du contenu.

Alors que l'alignement à gros grain s'appuie à la fois sur le marquage fort des paragraphes (*via* la mise en page) et faible des phrases (*via* la ponctuation), en corrélation avec le parallélisme présumé de la narration, l'alignement sous-phrastique se heurte immédiatement à la délimitation des unités, notamment lorsque le mot n'est pas physiquement marqué, ou bien lorsque la langue est agglutinante. À un grain sous-phrastique, on ne peut par ailleurs présumer une quelconque préservation de l'ordre des unités dans la phrase, sauf à se limiter à l'analyse de couples de langues proches. Pour pallier cette difficulté, le recours à un dictionnaire bilingue est souvent proposé, mais on peut regretter que les alignements produits ne soient alors que la mise en évidence de traductions déjà contenues dans le dictionnaire. On regrettera également que cette technique exclue l'analyse des langues faiblement dotées en matière de ressources linguistiques.

Enfin, on notera que l'alignement au niveau sous-phrastique est généralement effectué successivement à un alignement phrastique et que la qualité de ce dernier n'est donc pas sans effet sur la qualité des résultats du premier.

2. UNE APPROCHE ENDOGENE DE L'ALIGNEMENT SOUS-PHRASTIQUE

L'approche proposée ici ne prétend pas répondre à l'ensemble des limites précédemment énoncées. Elle concerne le repérage d'équivalents de traduction de taille variable, et notamment supérieure à un mot. La méthode est appliquée à un corpus parallèle de bi-textes anglais-grec. Nous n'adoptons pas de contraintes très rigides concernant la taille des unités linguistiques afin de tenir compte de la souplesse des langues et des divergences de traduction. Les liens d'alignement peuvent alors être établis à des niveaux variés – allant

de la phrase jusqu'au mot – en n'obéissant à d'autres contraintes qu'au nombre d'occurrences minimal. En ce qui concerne le prétraitement des textes d'entrée, ils ont été segmentés et alignés au niveau des phrases, mais ils ne portent aucune information morphosyntaxique et ne sont pas lemmatisés. Les divergences de flexion des mots isolés sont gérées automatiquement par un module intégré dans l'outil d'alignement, qui prend en compte les variants morphologiques. Ceci est très important surtout pour des langues comme le grec, qui est une langue hautement flexionnelle, avec une morphologie riche, et offre la possibilité de prendre en compte les formes fléchies différentes des mots.

2.1. Approche

Pour traiter ce problème, nous avons, dans la mesure du possible, essayé de nous tenir à une approche de type minimaliste, selon la méthode du groupe Syntaxe et Rhétorique de l'équipe ISLanD du GREYC. Nous n'utilisons donc pas de connaissances linguistiques particulières sur les langues à aligner.

En effet, beaucoup de langues sont peu dotées en matière de ressources libres pour le traitement automatique des langues, que ce soit par exemple en matière d'étiquetage morpho-syntaxique ou, plus généralement, d'analyse syntaxique de surface, ou bien tout simplement de ressources lexicales (dictionnaires électroniques etc.). De même, la possibilité de traitement de textes qui n'ont pas subi une longue chaîne de prétraitements est avantageuse vu la disponibilité de plus en plus grande de corpus bilingues.

Nous préférons donc des solutions fondées sur la recherche de motifs situés à des positions bien identifiées, combinée avec des indices de mise en forme matérielle (graisse, taille, couleur des caractères, centrage, justification etc.) ou bien typo-dispositionnelle (ponctuations, énumérations etc.).

2.2. Le corpus

Le corpus utilisé pour nos expériences est constitué de bi-textes alignés au grain phrase. Formellement, un bi-texte est un quadruplet <T1, T2, Fs, C> où T1 et T2 sont les deux textes, Fs est la fonction qui réduit T1 à un ensemble d'éléments Fs(T1) et T2 à un ensemble d'éléments Fs(T2), et C est un sous-ensemble du produit cartésien de Fs(T1) x Fs(T2) (Harris, 1988). Concrètement, un bi-texte aligné est constitué d'un ensemble d'alignements, c'est-à-dire un ensemble de correspondances entre phrases de la langue source et phrases de la langue cible. Ces alignements contiennent de 0 à 2 phrases par langue. Par exemple, un alignement « 2-1 » met en correspondance 2 phrases du texte 1 avec 1 phrase du texte 2, et un alignement « 1-0 » indique qu'une phrase du texte 1 n'a pas de correspondance dans le texte 2.

Le corpus est composé de 93 bi-textes anglais-grec (un peu plus de 200 000 mots). Il s'agit de résumés d'articles scientifiques parus dans des journaux internationaux de médecine (en anglais ou en grec) ainsi que leur traduction. Une partie du corpus est constituée de résumés d'articles parus dans le journal médical grec *Archives of Hellenic Medicine* de 1999 à 2004 (http://www.mednet.gr/archives/) et de leurs traductions en anglais. L'autre partie du corpus est constituée de résumés d'articles en anglais parus dans des journaux internationaux disponibles sur la base de données PubMed (www.pubmed.com). Les traductions de ces textes en grec ont paru dans des journaux publiés chez l'éditeur Scientific Publications : *Review of Risk Factors in Cardiology*, *Review of Diabetes*, *Review of Endocrinology – Metabolism*, *Review of Vaccines*, *Review of Paediatric Nutrition*. Il faut noter que même s'il s'agit de textes d'un domaine spécialisé

(médecine), il n'y a pas de grande homogénéité entre eux (surtout au niveau terminologique), étant donné qu'ils relèvent de sous-domaines variés (cardiologie, endocrinologie, pédiatrie, diabétologie, vaccinologie).

3. MÉTHODE D'ALIGNEMENT AUTOMATIQUE

La méthode de résolution se déroule en deux phases basées sur deux hypothèses sous-jacentes. La première porte sur les documents et la seconde sur l'ensemble du corpus.

3.1. Hypothèses

• **Hypothèse 1** : soit un bi-texte composé des textes T1 et T2. Si une séquence S1 est répétée dans T1 dans un certain nombre de phrases P1i, nous supposons qu'une séquence S2 correspondant à la traduction de S1 apparaîtra dans les phrases P2j de T2 où P2j sont les phrases alignées avec P1i.
• **Hypothèse 2** : soit un corpus de bi-textes composé des langues L1 et L2. Il n'y a pas de garantie pour qu'une séquence S1 répétée dans plusieurs textes de L1 ait une unique traduction dans les textes correspondants de la langue L2.

3.2. Phase 1 : Alignement intra-document

3.2.1. PRÉPARATION DU CORPUS L'algorithme prend en entrée le corpus de bi-textes alignés au niveau phrastique. Cet alignement a été obtenu automatiquement, par l'outil d'alignement de phrases du Système de Mémoire de Traduction TrAid (Triantafyllou et al., 2000). L'algorithme utilisé se base sur le modèle statistique de Gale & Church (1991a) qui tient compte de la longueur des phrases en nombre de caractères. Le résultat du processus d'alignement a été validé à la main. La segmentation en phrases et leur alignement ont été les seuls prétraitements du corpus.

Le corpus aligné est au format TMX[1], codage UTF-8. Chaque alignement est balisé par un élément XML, qui contient les segments de chaque langue mis en correspondance pendant le processus d'alignement qui, eux aussi, sont inclus entre des balises XML (Figure 1).

```
<tu>
    <prop type="Domain">ALIGNER: [med1_EN.txt]---[med1_EL.txt]</prop>
    <tuv lang="EN">
        <seg>Frequency and relevance of elevated calcitonin levels in patients with neoplastic and nonneoplastic
        thyroid disease and in healthy subjects.</seg>
    </tuv>
    <tuv lang="EL">
        <seg>Συχνότητα και σημασία των αυξημένων επιπέδων καλσιτονίνης σε ασθενείς με και χωρίς
        νεοπλασματική νόσο του θυρεοειδούς και σε υγιή άτομα</seg>
        1 Translation Memory eXchange (http://www.lisa.org/tmx)

    </tuv>
</tu>
```

Figure 1. La structure des fichiers TMX

La première étape consiste à écarter du traitement les alignements des bi-textes pour lesquels des segments d'une langue n'ont pas de correspondance dans l'autre (alignement « 1-0 », « 0-1 », « 2-0 » et « 0-2 »). On profite de cette vérification pour relever dans les deux parties du corpus, les index (offsets) de début et de fin de chaque segment.

3.2.2. IDENTIFICATION DES SÉQUENCES CANDIDATES À L'ALIGNEMENT Pour chacun des deux textes des bi-textes, nous calculons ensuite les séquences de mots répétées, ainsi que leur effectif (nombre d'occurrences). L'algorithme est paramétré par le nombre de mots minimal et maximal des séquences à mémoriser ainsi que par leur effectif minimal, mais ces paramètres ne sont utilisés qu'en phase de test. Par ailleurs, l'algorithme ne conserve pas les sous-séquences d'une séquence répétée si elles ont même effectif. Par exemple, si « heart disease » a même effectif que « coronary heart disease », nous ne retenons que la deuxième séquence. Par contre, si « disease » a un effectif supérieur à « coronary heart disease», nous retenons les deux.

Le calcul de l'effectif d'une séquence répétée s'accompagne du relevé des index de ses occurrences. On obtient donc à la fin du traitement une liste de séquences avec pour chacune son effectif, ainsi que la liste des index de ses occurrences dans le texte.

« healthy subjects » : index des 22 occurrences dans le texte 1 (44 segments)

285, 531, 984, 2860, 3993, 4281, 4386, 4646, 4781, 5122, 5242, 5382, 5473, 6615, 7370, 7422, 7690, 8856, 9019, 9459, 9553, 10025

3.2.3. PASSAGE À UNE REPRÉSENTATION VECTORIELLE DES SÉQUENCES Nous construisons ensuite un espace orthonormé pour explorer l'existence de relation de traduction entre séquences et définir les couples traductionnels internes à chaque bi-texte. Pour cela, nous notons pour chaque occurrence de séquence répétée les numéros où elle apparaît.

« healthy subjects » : liste des alignements où apparaît la séquence

2 , 4 , 7 , 13 , 15 , 16 , 16 , 17 , 17 , 18 , 19 , 20 , 21 , 30 , 34 , 34 , 35 , 39 , 39 , 41 , 41 , 42

Nous convertissons cette liste en un vecteur à *n* dimensions (où *n* correspond au nombre d'alignement du bi-texte). Chaque dimension contient le nombre d'occurrences de la séquence présent dans l'alignement.

« healthy subjects » : vecteur à 44 dimensions associé

12345678910111213141516171819…41424344010100100000101221 1…2100

3.2.4. ALIGNEMENT DES SÉQUENCES À L'AIDE DU COSINUS Nous distinguons dorénavant les deux langues du corpus, la langue L1 et la langue L2. Pour chaque séquence candidate à l'alignement dans la langue L1, nous explorons l'existence d'une relation de traduction entre elle et chacune des séquences de la langue L2 candidates à l'alignement. L'existence d'une relation de traduction entre deux séquences est estimée par le cosinus des vecteurs qui leur sont associés. Il est obtenu en divisant le produit scalaire des deux vecteurs par le produit de leurs normes :

$$\cos(x,y) = \frac{\sum_{i=1}^{22} x_i \times y_i}{\sqrt{\sum_{i=1}^{22} x_i^2} \times \sqrt{\sum_{i=1}^{22} y_i^2}}$$

Les séquences proposées pour l'alignement sont celles qui obtiennent le cosinus le plus grand. On ne propose pas d'alignement si le meilleur cosinus est inférieur à un seuil donné (on notera que ce cosinus n'est jamais négatif puisque les composantes des vecteurs sont toujours positives).

3.3. Phase 2 : Alignement inter-document

À l'issue de première phase de résolution à l'échelle des documents, nous disposons d'un certain nombre d'hypothèses de relation de traduction entre des segments d'une langue vers une autre. Le passage à l'échelle du corpus va nous permettre de confirmer ou d'infirmer les hypothèses précédemment posées.

Pour qu'une hypothèse résiste au filtrage au niveau du corpus elle doit respecter l'un des deux critères suivant :

1. avoir été extraite d'au moins deux bi-textes ;
2. si elle n'a été trouvée que dans un seul bi-texte alors les deux séquences doivent être strictement identiques ou bien avoir une fréquence supérieure à un seuil déterminé empiriquement (6 en pratique).

Les relations de traduction sont ensuite triées en fonction du nombre d'occurrences et du nombre de documents dont elles ont été extraites.

3.4. Évaluation

Nous avons réalisé une évaluation portant sur tous les bi-textes anglais-grec. Les résultats montrent que les couples traductionnels générés sont de très bonne qualité mais encore peu nombreux. 1 066 alignements ont été proposés parmi lesquels 60 sont erronés et 32 sont incomplets. Ainsi, le taux de précision est de 91 % en considérant les résultats incomplets comme erronés. Si l'on admet des résultats incomplets comme acceptables le taux de précision est de 94 %. Pour cette évaluation, les alignements considérés comme corrects sont ceux qui présentent une équivalence traductionnelle à la fois correcte et complète, c'est-à-dire sans mots supplémentaires ou manquants, notamment dans le cas de termes ou d'expressions mis en relation.

Pour une partie des alignements proposés, on constate un recouvrement partiel entre les séquences : il s'agit de termes complexes ou d'expressions de la langue source qui ont comme équivalent un terme complexe ou une expression dans la langue cible, mais dont seule une partie a été repérée par notre algorithme. On trouve réciproquement parfois dans la langue source une partie d'un terme complexe ou d'une expression qui est mise en équivalence avec un terme complet ou une expression complète dans la langue cible. Il est intéressant de noter que les correspondances incomplètes sont parfois accompagnées de correspondances complètes (p.ex. type – διαβήτη τύπου / type 2 diabetes – διαβήτη τύπου 2, as – καθώς και / as well as – καθώς και). Il s'agit là de sous-séquences d'une séquence répétée qui ont un effectif supérieur à celui de la séquence complète et qui sont ainsi retenues (voir section 3.2.2.). Ce phénomène s'explique en partie par le fait que l'on n'adopte pas de contraintes très rigides concernant la taille des unités linguistiques à aligner. Ce choix nous offre la possibilité de prendre en compte la flexibilité de la langue et de proposer des correspondances à des niveaux différents mais il se révèle parfois problématique.

Lorsqu'un terme complexe est suivi d'une abréviation dans un des textes, c'est souvent l'abréviation qui est mise en correspondance avec le terme complexe équivalent (tout ou une partie de ce dernier) et non le terme complexe (p.ex. DM / σακχαρώδη διαβήτη). Cette situation se rencontre lorsque l'abréviation est systématiquement utilisée à la place du terme complexe dans l'un des deux textes. Dans ces cas, on a considéré les résultats comme incomplets.

Nous ne disposons pas pour le moment de mesure précise du silence, mais il est non négligeable car la méthode ne considère comme candidates à l'alignement que les séquences répétées au sein du document. Les séquences qui ne sont jamais répétées au grain document n'ont par conséquent aucune chance d'aboutir à un alignement, et cela, même si ces séquences hapax au grain document apparaissent dans plusieurs documents du corpus. Ce choix permet cependant de contrôler l'explosion combinatoire.

La nature des alignements proposés est variée. Il ne s'agit pas seulement de mots pleins mais également de mots fonctionnels (articles, conjonctions, etc.) ainsi que de termes complexes ou d'expressions. Peu de verbes sont extraits, ce qui peut entre autre s'expliquer par le fait qu'au grain document la répétition lexicale concerne davantage le contenu nominal que le contenu verbal. Malgré la liberté de la méthode quant à la longueur des séquences appariées, on constate que les alignements proposés correspondent parfois au concept de *chunk* mais que bien souvent, ils ne constituent pas des unités de nature bien identifiée. On trouve des expressions ou bien plus simplement des portions de phrases sans statut bien établi.

Parmi les couples corrects, on trouve des couples de termes relatifs à la ligne éditoriale et à la structure des sources des documents : nom de la source, sous-titres de fiche résumé, dates (par exemple Αρχ Ελλ Ιατρ – Arch Hellen Med, ΑΠΟΤΕΛΕΣΜΑΤΑ – RESULTS ΣΥΜΠΕΡΑΣΜΑΤΑ – CONCLUSIONS), des couples de termes liés au domaine médical en général : des unités de mesures, des abréviations et des termes latins (p.ex. ml – ml, SH – SH, aeruginosa – aeruginosa, vitro – vitro) et des couples de termes liés à un sous-domaine de spécialité (p.ex. κορτιζόλης – cortisol, ομοκυστεΐνης – homocysteine, τεστοστερόνης – testosterone). On note également la présence de segments de longueur variable (σε συνδυασμό με – in combination with, Αρχ Ελλ Ιατρ, 19(6), Νοέμβριος-Δεκέμβριος 2002 – Arch Hellen Med, 19(6), November-December 2002).

Enfin, le traitement d'une grande partie des variantes morphologiques en grec et en anglais améliore beaucoup les résultats. Ainsi, on a parmi les résultats des correspondances comme : (power, ισχύ~) (εξέταση~ , examination) (factor~ , κινδύνου) (protein~ , πρωτεΐνη~)(memory, μνήμη~).

Les propositions d'alignement erronées sont peu nombreuses mais on constate que dans la quasi totalité des cas, l'alignement n'est extrait que d'un ou deux documents.

4. CONCLUSION

Nous avons proposé une méthode permettant d'extraire automatiquement des relations de traduction à partir d'un corpus bilingue. Un avantage de cette méthode est qu'elle n'utilise que très peu de connaissances linguistiques et qu'elle peut donc être appliquée à des langues ne disposant que de peu de ressources électroniques disponibles ou facilement accessibles. Il s'agit d'une méthode d'analyse endogène : les connaissances ne sont extraites que du corpus. Aucun dictionnaire n'est *a priori* exigé et les affixes grammaticaux sont inférés du corpus.

La méthode est multilingue et peut s'appliquer à toute langue. Pour des raisons d'efficacité, nous avons cependant intégré des connaissances sur les frontières des mots : existence du mot graphique et liste de ponctuations.

La méthode prend en entrée un corpus de bi-textes alignés au niveau phrastique. Elle produit des alignements sous-phrastiques de longueur variable. Bien qu'il soit de plus en plus aisé de trouver des corpus bilingues alignés au niveau phrastique, et cela pour des

couples de langues très variés, il est important de noter que la qualité de l'alignement phrastique fourni en entrée n'est pas sans influence sur la qualité des résultats générés par notre méthode. Lorsque la qualité de l'alignement phrastique chute, la qualité des alignements proposés reste stable mais les propositions sont alors moins nombreuses.

La méthode que nous utilisons considère différemment le grain document et le grain corpus. Ceci permet de maîtriser l'explosion combinatoire, d'accroître la confiance dans les résultats et d'obtenir plusieurs hypothèses pour une séquence unique en cas d'homographie. Ce résultat est cependant obtenu en échange d'une augmentation du silence, silence dont nous cherchons aujourd'hui une métrique d'évaluation.

Nos travaux s'orientent vers l'alignement de segments discontinus, dans une perspective multilingue et endogène, avec un regard particulier sur la maîtrise de la combinatoire. Par ailleurs, nous souhaitons intégrer un module de résolution de coréférences qui perturbent actuellement la détection (Giguet & Lucas, 2004).

5. RÉFÉRENCES

Altenberg Bengt & Sylviane Granger. 2002. 'Recent trends in cross-linguistic lexical studies' in Altenberg & Granger (eds.), *Lexis in Conrast, Corpus-based approaches*, Amstedam / Philadelphia : John Benjamins Publishing Company, 2002, p. 3-48.

Boutsis, S., & Piperidis, S. 1998. 'Aligning clauses in parallel texts' in *Third Conference on Empirical Methods in Natural Language Processing*, 2 June, Granada, Spain, 1998, p. 17-26.

Brown P., J. Lai & R. Mercer. 1991b. 'Aligning sentences in parallel corpora' in *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June, Berkley, California, 1991, p. 169-176.

Gale W.A. & K.W. Church. 1991a. 'Identifying word correspondences in parallel texts' in *Fourth DARPA Speech and Natural Language Workshop*, San Mateo, California : Morgan Kaufmann, 1991, p. 152-157.

Gale W.A. & K. W. Church. 1991b. 'A Program for Aligning Sentences in Bilingual Corpora' in *Proc. 29th Annual Meeting of the Association for Computational Linguistics*, 18-21 June, Berkley, California, 1991, p. 177-184.

Giguet E. & Lucas N. 2004. 'La détection automatique des citations et des locuteurs dans les textes informatifs' in López-Muñoz J. M., Marnette S., Rosier L. (eds.).*Le discours rapporté dans tous ses états : Question de frontières*, Paris : l'Harmattan, 2004, p. 410-418.

Harris B. 1998. 'Bi-text, a New Concept in Translation Theory', *Language Monthly*, (54), 1998, p. 8-10.

Isabelle Pierre & Susan Warwick-Armstrong. 1993. 'Les corpus bilingues : une nouvelle ressource pour le traducteur' in Bouillon, P. & Clas A. (eds.), *La Traductique : études et recherches de traduction par ordinateur*, Montréal : Les Presses de l'Université de Montréal, 1993, p. 288-306.

Kay M. & M. Röscheisen. 1993. 'Text-translation alignment', *Computational Linguistics*, March 1993, p. 121-142.

Kitamura & Matsumoto, 1996. 'Automatic extraction of word sequence correspondences in parallel corpora' in *Proc. 4th Workshop on Very Large Corpora*, Copenhagen, Denmark, 4 August 1996, p. 79-87.

Kupiec J. 1993. 'An algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora' in *Proc. of the 31st Annual Meeting of the Association of Computational Linguistics*, p. 23-30.

Papageorgiou Harris, Lambros Cranias & Stelios Piperidis. 1994. 'Automatic alignment in parallel corpora' in *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, 27-30 June, Las Cruses, New Mexico, 1994, p. 334-336.

Salkie Raphael. 2002. 'How can linguists profit from parallel corpora ?' in Lars Borin (ed.), *Parallel Corpora, Parallel Worlds : selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April, 1999, Amsterdam, New York : Rodopi, 2002, p. 93-109.

Simard, M., G. Foster, and P. Isabelle. 1992. 'Using cognates to align sentences in bilingual corpora' in *Proc. of TMI-92*, Montréal, Québec, 1992, p. 67-81.

Smadja F. 1992. 'How to compile a bilingual collocational lexicon automatically' in *Proc. of the AAAI-92 Workshop on Statistically -based NLP Techniques*, 1992, p. 65-71.

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. 1996. 'Translating Collocations for Bilingual Lexicons : A Statistical Approach', *Computational Linguistics*, March, 1996, p. 1-38.

Tiedemann Jörg. 1993, 'Combining clues for word alignment' in *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Hungary, April 2003, p. 339-346.

Triantafyllou Ioannis, Iason Demiros, Christos Malavazos & Stelios Piperidis. 2000. 'An alignment architecture for Translation Memory bootstrapping' in *Proc. of the MT 2000 : Machine Translation and Multilingual Applications in the New Millenium*, Exeter, United Kingdom, 20-22 November 2000, p. 3.1-3.8.